# Data Cleaning

January 27, 2026

# Today's plan

1. **Review: Computing Tools**
2. **What is data cleaning?**
3. **Jump-start with Lab 1 data cleaning**

# Review: Computing Tools

# Computing Tools

Details at: https://tiffanymtang.github.io/dsip-s26/

- We set up two **GitHub** repositories:
  - `dsip-s26`: for me to distribute course materials (lecture slides, **code**, etc) to you
    - You should only *pull* to retrieve information
  - `dsip`: **your** repository to do all your work in
    - You can *pull and push* to this repository
- We introduced **`renv`** and **`conda`** to create reproducible environments
- We introduced **quarto** to render reproducible documentation

# Basic Setup for Today (and most days in general)

1. Go to **dsip-s26** repository and git pull to retrieve materials for today
2. Copy lab1/ folder into your **dsip/** folder
    - Should be dsip/lab1/
3. Restore renv or conda environment so that you have all necessary packages
    - **R:**
        i. Open dsip/lab1/ project in Positron/VS Code/RStudio
        ii. `renv::restore()`
    - **Python:**
        i. Navigate to dsip/lab1/ folder in terminal
        ii. `conda-lock install --name dsip_lab1`
        iii. Activate conda environment: `conda activate dsip_lab1`

# Pushing changes to GitHub

We've completed our basic setup – A great time to pause and take a snapshot of our project.

But before doing so, if you check your git status,

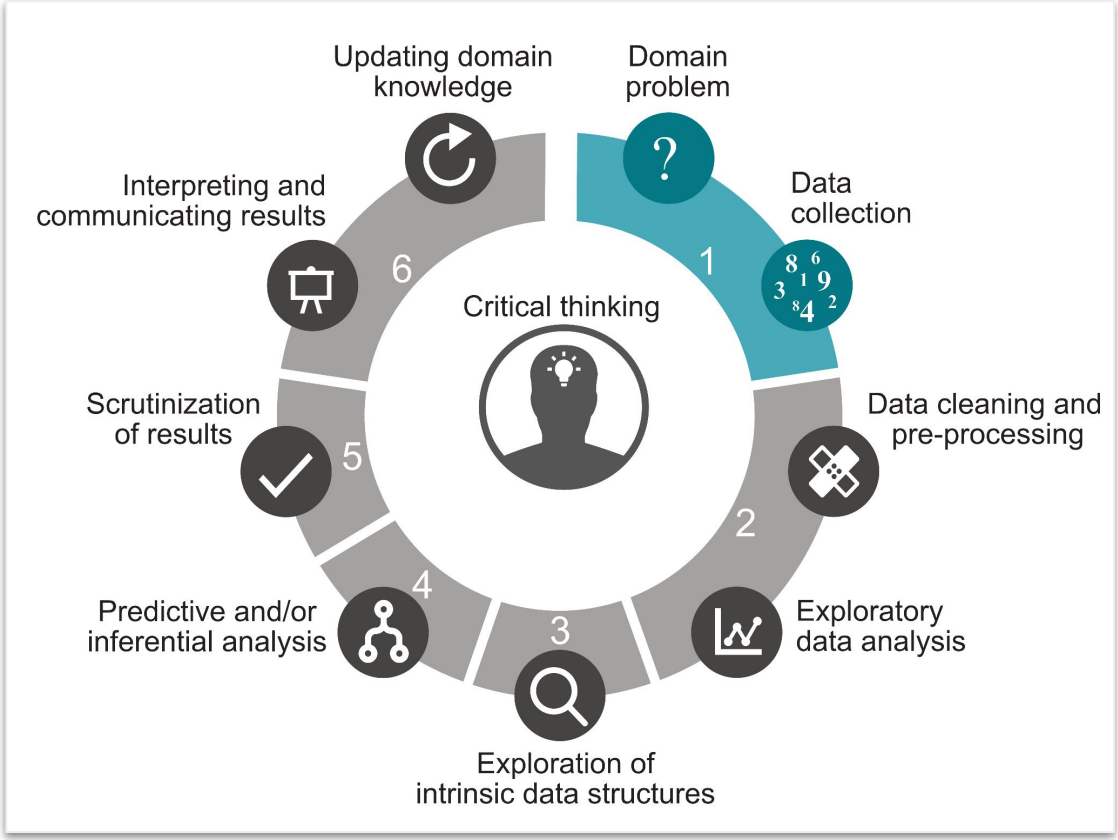
- You might notice a lot of "junk" files that aren't worth tracking (e.g., they are useless to collaborators and/or they change too frequently)
- Add these files to your .gitignore, e.g.,

```
*.DS_Store
*/data/*
*__pycache__*
*.ipynb_checkpoints*
```

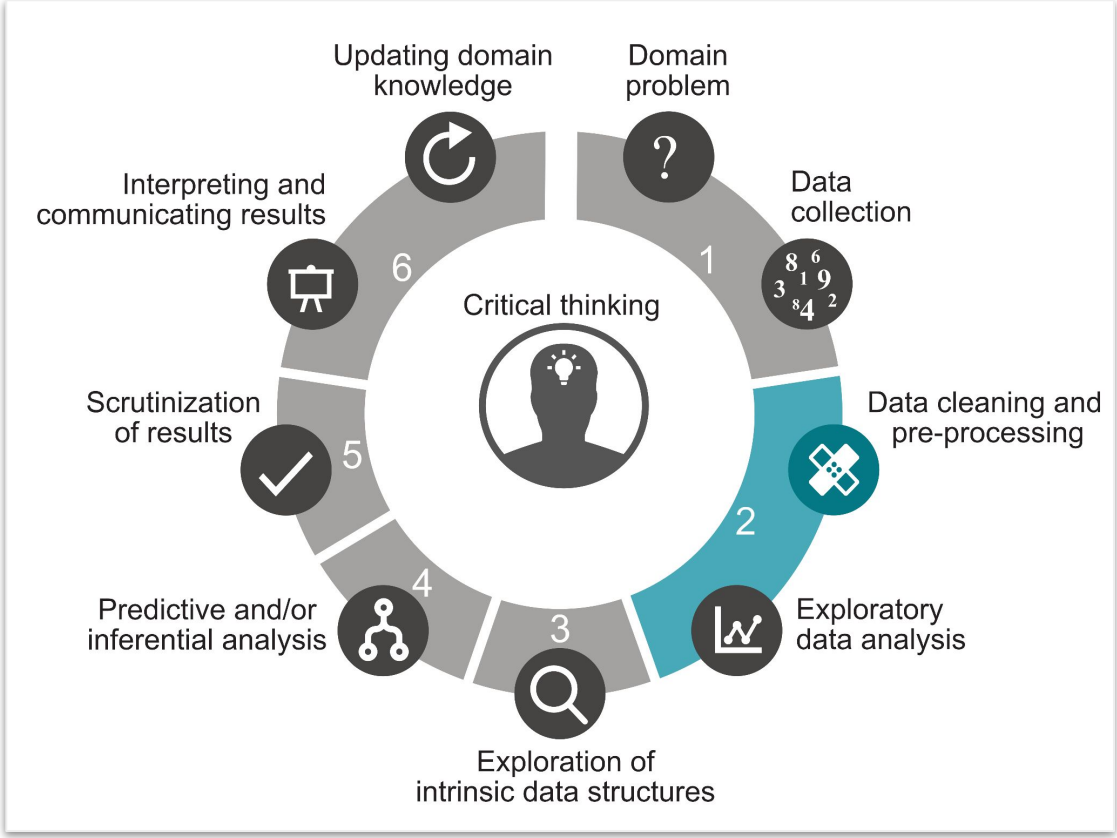
Details:
https://tiffanymtang.github.io/dsip-s26/course_materials/tools/01_git_github.html#gitignore

# Data Cleaning

# So far…


Updating domain knowledge
Domain problem
Interpreting and communicating results
Data collection
6
1
Critical thinking
Scrutinization of results
Data cleaning and pre-processing
5
2
4
Predictive and/or inferential analysis
3
Exploratory data analysis
Exploration of intrinsic data structures

# What's next?


Domain problem
1
Data collection
2
Data cleaning and pre-processing
3
Exploration of intrinsic data structures
4
Predictive and/or inferential analysis
5
Scrutinization of results
6
Interpreting and communicating results
Updating domain knowledge
Critical thinking
Exploratory data analysis

# What is data cleaning?


"The data cleaning procedure will involve learning about the data collection process, obtaining domain knowledge, examining the data, asking and answering questions, checking assumptions, and identifying and documenting any issues and inconsistencies..."

- Veridical Data Science (VDS) Textbook (ch 4)

+ It's an **iterative** and **interactive** process. You are like a detective.
+ **Poll:** What are common issues that you might look for when data cleaning?
+ **Poll:** What are some things you can do to identify these data issues?

# How to *identify* issues with the data

- Read all documentation
- Look at summary statistics, NAs, duplicates
  - `skimr::skim()` in R
- **Plot, plot, plot, and iterate**
  - Scatter plots, histograms, density plots, heatmaps, ...
- Use domain knowledge/common sense
  - Ask lots of questions to the data? Do the answers make sense?

# How to *fix* the identified issues

- **It depends** on your domain problem/goal and data collection
- Whatever you do, **DOCUMENT AND JUSTIFY YOUR DECISIONS**
  (ideally using your knowledge of the domain problem or data collection process)
  - Do NOT simply delete data because it is >3 SDs from the mean and is often defined as an "outlier" in traditional statistics textbooks
    - If you delete data, you should have a reason for deleting this data
  - For example: "We omitted data from all counties with median age > 100 because this is most likely not possible in reality and a result of measurement error"

# A quick note on missing values

## Types of Missingness

- **Missing Completely At Random** (MCAR): missingness unrelated to anything
  - Easy case
- **Missing At Random** (MAR): depends on observed variables *X*
  - Medium case
- **Missing Not At Random** (MNAR): depends on unobserved values themselves
  - Very hard case
  - Typically want to do a sensitivity analysis (try out a diverse array of procedures to handle missingness and see how this affects the results)

The further you move from MCAR → MNAR, the more careful you need to be.

## Document and be transparent!

Q: What are some ways we can deal with missing data?

# How to deal with missing data

- **Deletion Methods**
  - **Drop observations** (rows) with missing values: OK if MCAR
    (e.g., if a few observations have high rates of missingness)
  - **Drop features** (columns) with missing values: OK if MCAR
    (e.g., if a few features have high rates of missingness)
- **Imputation Methods**
  - Impute missing values with **mean/median/mode** feature value: OK if MCAR
  - **Regression Imputation:** use regression model (e.g., random forest) to predict missing values from other variables ($x_j$ ~ all other $x$'s); OK if MAR
  - **Multiple imputation:** create multiple imputed datasets, analyze each, and pool results
- And many other methods…

# Jump-start with cleaning redwood tree data

# Lab 1: A Macroscope in the Redwoods [Tolle et al. (2005)]

- **Coastal redwood trees:**
  very tall, very old
- 44-day study in Sonoma, California
  (April 27, 2004 5:10pm - June 10, 2004 2pm)

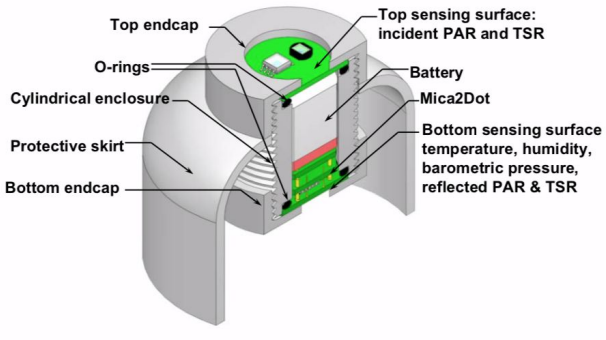



Figure 2: Sensor node and packaging

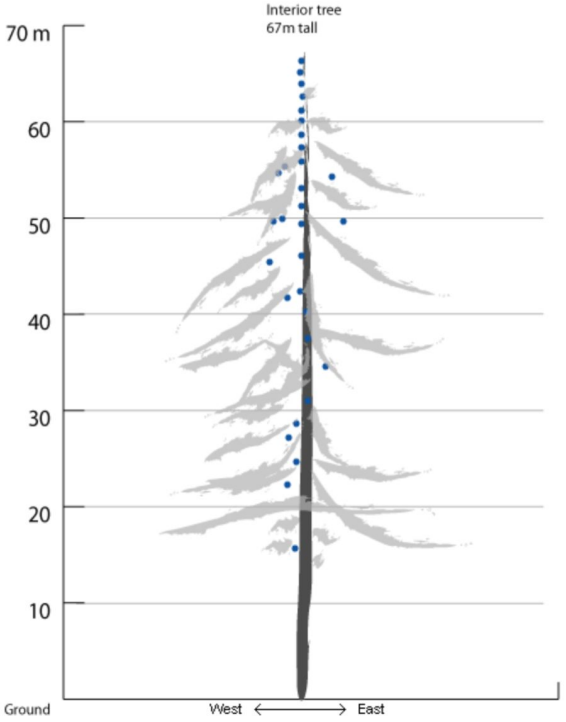
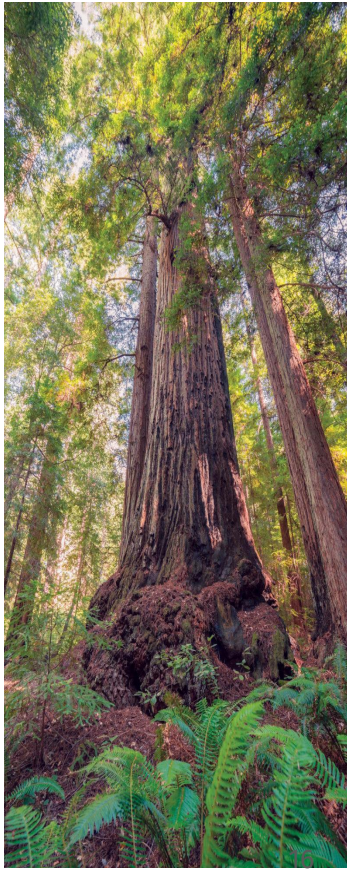



Figure 1: The placement of nodes within the tree



# Lab 1 Clarifications

- Problem formulation:
  - You are not expected to come up with a groundbreaking scientific question
  - Descriptive questions are ok, no “modeling” is necessary
    - End goal: exploratory data analysis/visualizations
  - The connection between your analysis and domain problem should be reasonable, not bulletproof
  - Ex: How much light reaches the bottom of a redwood tree?
    - Understanding the amount of light available gives us insight into the types of plants that are able to grow in that environment
  - 1-3 sentences per question in the problem formulation part should suffice
  - It’s ok to ask questions. This is part of collaboration.
- Do not wait until the last minute to work on this. Chip away at it; don’t binge.

# Getting started with the redwood lab

Available templates on the course repository (`dsip-s26`)

- **Reproducible report**: you can find quarto templates in `notebooks/`
  - In the end, this report should only contain your "filtered" code, not everything you ever did and or looked at in this lab.
- Some **loading** and **cleaning** functions have already been populated in `R/` and `python/` folders
  - Some functions have scaffolding but not yet filled in; this is your to-do

For today's class, you can use **notebooks/lab1_R.qmd** (or _python.qmd) or you can create your own exploratory "scratch" notebook

# Your tasks for today

1. **Load in the data**
   a. Epoch/dates and redwood datasets have already been filled out for you.
   b. You need to fill out the `load_mote_location_data()` in the load.R/load.py file.
      - The output of this function should be a data frame (or tibble) with 80 rows and 5 columns (column names: "ID", "Height", "Direc", "Dist", "Tree")

2. Look and "play" around with the data in order to:
   a. Try to **identify as many issues or oddities** with the data as you can. *Hint:* there are many!!
   b. Also **think** about how you might address these issues and clean the data. Jot down these ideas, but no need to take the time to implement it *yet*.
      - *Time permitting*: start implementing your ideas, but prioritize identifying the issues over fixing them.

# Recap + Next Time

**Recap**

- **Data cleaning** is a highly iterative process.
- My two cents:
  - Don’t be afraid to ask lots of questions. Better to ask than to assume (more likely than not, incorrectly)
  - Read all documentation

**Next Time**

- More hands-on practice with data cleaning + exploratory data analysis [chapters 4 and 5 from VDS textbook]